Simplifying optimal strategies in stochastic games

SIAM J. CONTROL OPTIM.
Vol. 36, No. 4, pp. 1331–1347, July 1998

012

SIMPLIFYING OPTIMAL STRATEGIES IN STOCHASTIC GAMES*

J. FLESCH†, F. THUIJSMAN†, AND O. J. VRIEZE†

Abstract. We deal with zero-sum limiting average stochastic games. We show that the existence of arbitrary optimal strategies implies the existence of stationary ε-optimal strategies, for all $\varepsilon > 0$, and the existence of Markov optimal strategies. We present such a construction for which we do not even need to know these optimal strategies. Furthermore, an example demonstrates that the existence of stationary optimal strategies is not implied by the existence of optimal strategies, so the result is sharp.

More generally, one can evaluate a strategy π for the maximizing player, player 1, by the reward $\phi_s(\pi)$ that π guarantees to him when starting in state s. A strategy π is called nonimproving if $\phi_s(\pi) \geq \phi_s(\pi[h])$ for all s and for all finite histories h with final state s, where $\pi[h]$ is the strategy π conditional on the history h. Using the evaluation ϕ, we may define the relation "ε-better" between strategies. A strategy π^1 is called ε-better than π^2 if $\phi_s(\pi^1) \geq \phi_s(\pi^2) - \varepsilon$ for all s. We show that for any nonimproving strategy π, for all $\varepsilon > 0$, there exists an ε-better stationary strategy and a (0-)better Markov strategy as well. Since all optimal strategies are nonimproving, this result can be regarded as a generalization of the above result for optimal strategies.

Finally, we briefly discuss some other extensions. Among others, we indicate possible simplifications of strategies that are only optimal for particular initial states by "almost stationary" ε-optimal strategies, for all $\varepsilon > 0$, and by "almost Markov" optimal strategies. We also discuss the validity of the above results for other reward functions. Several examples clarify these issues.

Key words. stochastic games, limiting average rewards, optimality, Markov strategies, stationary strategies

AMS subject classifications. 90D15, 90D20, 90D05

PII. S0363012996311940

1. Introduction.

A zero-sum stochastic game Γ can be described by a state space $S := \{1, \ldots, z\}$ and a corresponding collection $\{M_1, \ldots, M_z\}$ of matrices, where matrix M_s has size $m_s^1 \times m_s^2$ and, for $i_s \in I_s := \{1, \ldots, m_s^1\}$ and $j_s \in J_s := \{1, \ldots, m_s^2\}$, entry (i_s, j_s) of M_s consists of a payoff $r(s, i_s, j_s) \in \mathbb{R}$ and a probability vector $p(s, i_s, j_s) = (p(1|s, i_s, j_s), \ldots, p(z|s, i_s, j_s))$. The elements of S are called states and for each state $s \in S$ the elements of I_s and J_s are called the actions of player 1 and player 2 in state s. The game is to be played at stages in $\mathbb{N}$ in the following way. The play starts at stage 1 in an initial state, say, in state $s^1 \in S$, where, simultaneously and independently, both players are to choose an action: player 1 chooses an $i_{s^1}^1 \in I_{s^1}$, while player 2 chooses a $j_{s^1}^1 \in J_{s^1}$. These choices induce an immediate payoff $r(s^1, i_{s^1}^1, j_{s^1}^1)$ from player 2 to player 1. Next, the play moves to a new state according to the probability vector $p(s^1, i_{s^1}^1, j_{s^1}^1)$, say, to state s^2. At stage 2 new actions $i_{s^2}^2 \in I_{s^2}$ and $j_{s^2}^2 \in J_{s^2}$ are to be chosen by the players in state s^2. Then player 1 receives payoff $r(s^2, i_{s^2}^2, j_{s^2}^2)$ from player 2 and the play moves to some state s^3 according to the probability vector $p(s^2, i_{s^2}^2, j_{s^2}^2)$, and so on.

The sequence $(s^1, i_{s^1}^1, j_{s^1}^1; \ldots; s^{n-1}, i_{s^{n-1}}^{n-1}, j_{s^{n-1}}^{n-1}; s^n)$ is called the history up to stage n. The players are assumed to have complete information and perfect recall.

A mixed action for a player in state s is a probability distribution on the set of his actions in state s. Mixed actions in state s will be denoted by x_s for player 1 and

*Received by the editors November 11, 1996; accepted for publication (in revised form) June 5, 1997; published electronically May 27, 1998.
http://www.siam.org/journals/sicon/36-4/31194.html
†Department of Mathematics, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, the Netherlands (flesch@math.unimaas.nl, frank@math.unimaas.nl, oj.vrieze@math.unimaas.nl).

by y_s for player 2, and the sets of mixed actions in state s by X_s and Y_s, respectively. A strategy is a decision rule that prescribes a mixed action for any finite history of the play. Such general strategies, so-called behavior strategies, will be denoted by π for player 1 and by σ for player 2, and $\pi(h)$ and $\sigma(h)$ will denote the mixed actions for history h. We use the notations Π and Σ, respectively, for the behavior strategy spaces of players 1 and 2. If for all finite histories, the mixed actions prescribed by a strategy only depend on the current stage and state, then the strategy is called Markov, while if they only depend on the state then the strategy is called stationary. Thus the stationary strategy spaces are $X := \times_{s\in S} X_s$ for player 1 and $Y := \times_{s\in S} Y_s$ for player 2, while the Markov strategy spaces are $F := \times_{n\in\mathbb{N}} X$ for player 1 and $G := \times_{n\in\mathbb{N}} Y$ for player 2. We will use the respective notations x and y for stationary strategies and f and g for Markov strategies for players 1 and 2. A stationary strategy is called pure if, for each state, it specifies one "pure" action to be used. Hence the spaces of pure stationary strategies are $I := \times_{s\in S} I_s$ for player 1 and $J := \times_{s\in S} J_s$ for player 2. Pure stationary strategies will be denoted by i and j, respectively.

Let H denote the set of finite histories, $H(\alpha,\omega)$ the set of finite histories with initial state α and final state ω, $H(\alpha,\cdot)$ the set of finite histories with initial state α, and $H(\cdot,\omega)$ the set of finite histories with final state ω. For any strategy π and for any given history $h \in H(\cdot,\omega)$, we can define the strategy $\pi[h]$ which prescribes a mixed action $\pi[h](\bar{h})$ to each history $\bar{h} \in H(\omega,\cdot)$ as if h had happened before $\bar{h}$, i.e., $\pi[h](\bar{h}) = \pi(h\bar{h})$, where $h\bar{h}$ is the history consisting of h concatenated with $\bar{h}$.

Payoffs and transition probabilities can be naturally extended to mixed actions as well. For $x_s \in X_s$ and $y_s \in Y_s$ let

$$r(s,x_s,y_s) := \sum_{i_s\in I_s,\, j_s\in J_s} x_s(i_s)\, y_s(j_s)\cdot r(s,i_s,j_s),$$

$$p(t|s,x_s,y_s) := \sum_{i_s\in I_s,\, j_s\in J_s} x_s(i_s)\, y_s(j_s)\cdot p(t|s,i_s,j_s).$$

For $x \in X,\ y \in Y$ we will also use the vector notation

$$r(x,y) := (r(s,x_s,y_s))_{s\in S}.$$

A pair of strategies (π,σ) with an initial state $s \in S$ determines a stochastic process on the payoffs. The sequences of payoffs are evaluated by the limiting average reward and by the β-discounted reward, $\beta \in (0,1)$, given by

$$\gamma(s,\pi,\sigma) := \liminf_{N\to\infty} \mathbb{E}_{s\pi\sigma}\left(\frac{1}{N}\sum_{n=1}^{N} r_n\right) = \liminf_{N\to\infty} \mathbb{E}_{s\pi\sigma}\left(R_N\right),$$

$$\gamma^{\beta}(s,\pi,\sigma) := \mathbb{E}_{s\pi\sigma}\left((1-\beta)\sum_{n=1}^{\infty}\beta^{n-1} r_n\right),$$

where r_n is the random variable for the payoff at stage $n \in \mathbb{N}$, and R_N for the average payoff up to stage N. We also use the vector notations

$$\gamma(\pi,\sigma) := (\gamma(s,\pi,\sigma))_{s\in S}, \qquad \gamma^{\beta}(\pi,\sigma) := \left(\gamma^{\beta}(s,\pi,\sigma)\right)_{s\in S}.$$

A pair of stationary strategies (x, y) determines a Markov chain with transition matrix P_{xy} on S, where entry (s,t) of P_{xy} is $p(t|s, x_s, y_s)$. With respect to this Markov chain, we can speak of transient and recurrent states (a state is called recurrent if, when starting there, it will be visited infinitely often with probability 1; otherwise the state is called transient). We can group the recurrent states into minimal closed sets, and into so-called ergodic sets (an ergodic set is a collection E of recurrent states with the property that, when starting in one of the states in E, all states in E will be visited and the play will remain in E with probability 1). Let

$$Q_{xy} := \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} (P_{xy})^n;$$

the limit is known to exist (cf. Doob [1953, Theorem 2.1, p. 175]). Entry (s,t) of the stochastic matrix Q_{xy}, denoted by $q(t|s, x, y)$, is the expected average number of stages the process is in state t when starting in s. The matrix Q_{xy} has the well-known properties (cf. Doob [1953]) that

$$Q_{xy} = Q_{xy}\, P_{xy} = P_{xy}\, Q_{xy}, \qquad Q_{xy}^2 = Q_{xy}. \tag{1.1}$$

By its definition, for the limiting average reward we have

$$\gamma(x, y) = Q_{xy}\, r(x, y), \tag{1.2}$$

hence by (1.1) we also obtain

$$\gamma(x, y) = Q_{xy}\, r(x, y) = Q_{xy}^2\, r(x, y) = Q_{xy}\, \gamma(x, y). \tag{1.3}$$

Against a fixed stationary strategy y there always exists a pure stationary best reply i of player 1 (cf. Hordijk, Vrieze, and Wanrooij [1983]); i.e.,

$$\gamma(i, y) \geq \gamma(\pi, y) \qquad \forall \pi.$$

Obviously a similar statement holds for the best replies of player 2.

For the limiting average reward, Mertens and Neyman [1981] showed that

$$\sup_{\pi} \inf_{\sigma} \gamma(s, \pi, \sigma) = \inf_{\sigma} \sup_{\pi} \gamma(s, \pi, \sigma) =: v_s \qquad \forall s \in S. \tag{1.4}$$

Here $v := (v_s)_{s\in S}$ is called the limiting average value and v is known to satisfy the following equations:

$$v_s = \mathrm{Val}\,(A_s) \qquad \forall s \in S, \tag{1.5}$$

where

$$A_s := \left[\sum_{t\in S} p(t|s, i_s, j_s)\, v_t\right]_{i_s\in I_s, j_s\in J_s} \tag{1.6}$$

and Val stands for the matrix game value. The sets of optimal mixed actions in A_s, for any $s \in S$, are nonempty polytopes. A strategy π of player 1 is called optimal for initial state $s \in S$ if

$$\gamma(s, \pi, \sigma) \geq v_s \quad \forall \sigma \in \Sigma,$$

and ε-optimal for initial state $s \in S$, $\varepsilon > 0$ if

$$\gamma(s, \pi, \sigma) \geq v_s - \varepsilon \quad \forall \sigma \in \Sigma.$$

If a strategy of player 1 is optimal or ε-optimal for all initial states in S, then the strategy is called optimal or ε-optimal, respectively. Optimality for strategies of player 2 is analogously defined. Although for all $\varepsilon > 0$, by (1.4), there exist ε-optimal strategies for both players, the famous example of Gillette [1957], the Big Match, examined by Blackwell and Ferguson [1968], demonstrates that in general the players need not have optimal strategies, and for achieving ε-optimality, behavior strategies are indispensable.

For the β-discounted reward, $\beta \in (0, 1)$, using a fixed-point argument, Shapley [1953] showed that

$$\sup_{\pi} \inf_{\sigma} \gamma^{\beta}(s, \pi, \sigma) = \inf_{\sigma} \sup_{\pi} \gamma^{\beta}(s, \pi, \sigma) =: v_s^{\beta} \qquad \forall s \in S.$$

Here $v^{\beta} := (v_s^{\beta})_{s \in S}$ is called the β-discounted value. Optimality can be similarly defined as for the limiting average reward. Stationary β-discounted optimal strategies always exist, and x is β-discounted optimal if and only if

$$v_s^{\beta} \leq (1 - \beta)\, r(s, x_s, y_s) + \beta \sum_{t \in S} p(t|s, x_s, y_s)\, v_t^{\beta} \qquad \forall y_s \in Y_s,\ \forall s \in S.$$

We will also make use of the N-stage game Γ^N, $N \in \mathbb{N}$, which is played up to stage N and where the reward is defined by the expected average payoff up to stage N. The N-stage game Γ^N, $N \in \mathbb{N}$, is known to have a value v^N, and both players have N-stage Markov optimal strategies. Bewley and Kohlberg [1976] showed, using Puiseux series, that both $\lim_{\beta \uparrow 1} v^{\beta}$ and $\lim_{N \to \infty} v^N$ exist and

$$\lim_{\beta \uparrow 1} v^{\beta} = \lim_{N \to \infty} v^N,$$

while Mertens and Neyman [1981] proved that the limiting average value is equal to the limit of the β-discounted values, i.e.,

$$v = \lim_{\beta \uparrow 1} v^{\beta}.$$

Although both the β-discounted value and the limiting average value exist, they cannot usually be easily calculated. In general, only iterative algorithms are available. We refer to Raghavan and Filar [1991] for a survey on algorithms.

We will often deal with specific restricted games derived from Γ. Assume that $S' \subset S$ is a nonempty set of states and $X_s' \subset X_s$, $Y_s' \subset Y_s$ are nonempty polytopes for all $s \in S'$. If all pairs of mixed actions in $X_s' \times Y_s'$, for all $s \in S'$, only induce transitions to states in S', then we may define a restricted game Γ', derived from Γ, where the state space is S' and the players are restricted to use strategies that only prescribe mixed actions in X_s' and Y_s' if the play is in state $s \in S'$. Let $\Pi' \subset \Pi$ and $\Sigma' \subset \Sigma$ denote the sets of these strategies. Clearly, the stationary strategy spaces in Γ' are $X' := \times_{s \in S'} X_s'$ and $Y' := \times_{s \in S'} Y_s'$. For the restricted game Γ', with respect to the β-discounted reward, $\beta \in (0, 1)$, similar results can be shown by using a fixed-point argument as for the original game Γ. Thus

$$\sup_{\pi \in \Pi'} \inf_{\sigma \in \Sigma'} \gamma^{\beta}(s, \pi, \sigma) = \inf_{\sigma \in \Sigma'} \sup_{\pi \in \Pi'} \gamma^{\beta}(s, \pi, \sigma) =: v_s'^{\beta} \qquad \forall s \in S'.$$

Here $v'^{\beta} := (v'^{\beta}_s)_{s\in S'}$ is called the β-discounted value for Γ'. Stationary β-discounted optimal strategies in Γ' always exist and $x \in X'$ is β-discounted optimal if and only if

$$v'^{\beta}_s \le (1-\beta)\, r(s, x_s, y_s) + \beta \sum_{t\in S'} p(t|s, x_s, y_s)\, v'^{\beta}_t \qquad \forall y_s \in Y'_s,\ \forall s \in S'. \tag{1.7}$$

The results of Bewley and Kohlberg [1976] apply for Γ' as well, so $\lim_{\beta\uparrow 1} v'^{\beta}$ and $\lim_{N\to\infty} v'^{N}$ exist and

$$v'^{1} := \lim_{\beta\uparrow 1} v'^{\beta} = \lim_{N\to\infty} v'^{N}. \tag{1.8}$$

Note that we do not claim that v'^{1} is the limiting average value of Γ', for even though the players only observe pure actions, these do not correspond one-to-one to extreme points of the restricted spaces of mixed actions. However, one can show, by using an appropriate sequence of discount factors as in Mertens and Neyman [1981], that, against any fixed strategy in Π', for any $\varepsilon > 0$ player 2 can make sure that player 1's limiting average reward is at most $v'^{1} + \varepsilon$; i.e.,

$$\sup_{\pi\in\Pi'} \inf_{\sigma\in\Sigma'} \gamma(s, \pi, \sigma) \le v'^{1}_s \qquad \forall s \in S'. \tag{1.9}$$

From now on, when we speak of rewards, values, or optimal strategies, we will always have the limiting average reward in mind, unless mentioned otherwise.

The organization of the paper is as follows. In section 2 we will deal with optimal strategies. We show that the existence of arbitrary optimal strategies implies the existence of stationary ε-optimal strategies, for all $\varepsilon > 0$, and the existence of Markov optimal strategies. We give such a construction for which we do not even need to know any optimal strategy. This remarkable result should not only be regarded as a simplification of optimal strategies, but also as a sufficient condition for the existence of stationary ε-optimal strategies or Markov optimal strategies. For many classes of stochastic games, where on the payoff or transition structures special conditions are imposed, stationary ε-optimal strategies exist, for all $\varepsilon > 0$, while about sufficient conditions for the existence of Markov optimal strategies, comparatively little is known. Here, instead of providing such structural conditions, the existence of optimal strategies will be proven to be sufficient. Moreover, an example will be provided to show that the existence of stationary optimal strategies is not implied by the existence of optimal strategies, so the result is sharp.

In section 3 we show that simplification of strategies can also be employed for a class of strategies, containing the optimal ones, in view of the rewards they guarantee. For this purpose we will evaluate a strategy π by the reward $\phi_s(\pi)$ that π guarantees when starting in state $s \in S$. A strategy π is called "nonimproving" if $\phi_s(\pi) \ge \phi_s(\pi[h])$ for all s and for all finite histories h with final state s, where $\pi[h]$ is the strategy π conditional on the history h, as defined above. Intuitively, a nonimproving strategy, for any state, cannot guarantee a larger reward conditional on any past history than initially. Using the evaluation ϕ, we may naturally define the relation "ε-better" between strategies. A strategy π^1 is called ε-better than π^2 if $\phi_s(\pi^1) \ge \phi_s(\pi^2) - \varepsilon$ for all $s \in S$. We show that for any nonimproving strategy π, for all $\varepsilon > 0$, there exists an ε-better stationary strategy and a (0-)better Markov strategy as well. Optimal strategies are clearly nonimproving, since they guarantee the value and more cannot be guaranteed; hence this result implies the above result for optimal strategies.

In section 4 we briefly discuss some extensions of the above results. We indicate possible simplifications of strategies that are only optimal for particular initial

states by "almost stationary" ε-optimal strategies and by "almost Markov" optimal strategies. We also discuss the validity of the results when other rewards are used to evaluate the long-term average payoffs. Some remarks concerning the proofs and some consequences are mentioned.

2. Optimal strategies. In this section we show the following result.

THEOREM 2.1. *If player* 1 *has an optimal strategy then, for all* $\varepsilon > 0$*, player* 1 *has stationary* ε*-optimal strategies and Markov optimal strategies as well.*

The proof will be constructive. We present such a construction for which we do not even need to know the optimal strategy.

For $s \in S$ let

$$X_s^* := \left\{ x_s \in X_s | \sum_{t \in S} p(t|s, x_s, y_s)\, v_t \geq v_s \ \ \forall y_s \in Y_s \right\}, \qquad X^* := \times_{s \in S} X_s^*,$$

so X_s^* is the set of optimal mixed actions for player 1 in the matrix game A_s (cf. (1.6)). The sets X_s^* are nonempty polytopes. For $s \in S$ let

$$Y_s^* := \left\{ y_s \in Y_s | \sum_{t \in S} p(t|s, x_s, y_s)\, v_t = v_s \ \ \forall x_s \in X_s^* \right\}, \qquad Y^* := \times_{s \in S} Y_s^*;$$

the sets Y_s^*, called the equalizers in the corresponding matrix games, are nonempty polytopes (in fact, by (1.5) all optimal mixed actions of player 2 in A_s belong to Y_s^*). Note the asimilarity in the definitions of X_s^* and Y_s^*, $s \in S$. It is easy to verify that, for any $s \in S$, there exists a $J_s^* \subset J_s$ such that $Y_s^* = \text{conv}(J_s^*)$, where conv stands for the convex hull of a set. Let

$$J^* := \times_{s \in S} J_s^*.$$

As described in the Introduction, we may define a restricted game Γ^*, derived from Γ, where the state space is S and the players are restricted to use strategies that only prescribe mixed actions in X_s^* and Y_s^* if the play is in state $s \in S$. The sets of these strategies are denoted by Π^* and Σ^*. Let $v^{*\beta}$ denote the β-discounted value for Γ^*, and let $v^{*1} := \lim_{\beta \uparrow 1} v^{*\beta}$.

By the finiteness of the state and action spaces there exists a countable subset of discount factors $\mathcal{B} \subset (0,1)$ such that 1 is a limit point of $\mathcal{B}$ and there are stationary β-discounted optimal strategies $x^\beta \in X^*$ in the restricted game Γ^* such that the sets $\{i_s \in I_s | x_s^\beta(i_s) > 0\}$, $s \in S$, are independent of $\beta \in \mathcal{B}$. In the sequel each time that we are dealing with discount factors, discounted optimal strategies, or limits when the discount factors converge to 1, we will have such a subset of discount factors $\mathcal{B}$ in mind.

The following lemma clarifies why the sets X^* and Y^* play an important role when player 1 has an optimal strategy in the original game Γ. This lemma states that if π is an optimal strategy for player 1 in Γ then, for any history with a positive occurrence probability with respect to (π, σ) for some $\sigma \in \Sigma^*$, the strategy π prescribes a mixed action belonging to X^*. In other words, if player 2 uses a strategy $\sigma \in \Sigma^*$ then the optimal strategy π will behave as a strategy in Π^*.

LEMMA 2.2. *Let* $\pi \in \Pi$ *be an optimal strategy for player* 1 *in the game* Γ*. Then for all* $h \in H(\alpha, \omega)$*, for any* $\alpha, \omega \in S$*, we have* $\pi(h) \in X_\omega^*$ *if* $\mathbb{P}_{\alpha\pi\sigma}(h) > 0$ *for some* $\sigma \in \Sigma^*$*. Here* $\mathbb{P}_{\alpha\pi\sigma}(h)$ *denotes the probability that the finite history* h *occurs when the strategies* π *and* σ *are played and the initial state is* α*.*

Proof. Suppose the opposite. Then there exists a shortest history $\bar{h}^n \in H(\alpha,\omega)$, say, with length n, for some $\alpha,\omega \in S$, and a $\sigma \in \Sigma^*$ with $\mathbb{P}_{\alpha\pi\sigma}(\bar{h}^n) > 0$ such that $\pi(\bar{h}^n) \notin X^*_\omega$. Since $\pi(\bar{h}^n) \notin X^*_\omega$ there exists a $j_\omega \in J_\omega$ such that

$$\tau := v_\omega - \sum_{s\in S} p(s|\omega,\pi(\bar{h}^n),j_\omega)\, v_s > 0.$$

Let $s^1 := \alpha$, let s^k, $k \geq 2$, denote the random variable for the state at stage k, and let h^k denote the history up to stage $k \in \mathbb{N}$. Let

$$\delta \in \left(0, \mathbb{P}_{s^1\pi\sigma}(\bar{h}^n)\cdot\tau\right).$$

Let $\sigma^\delta \in \Sigma$ be the strategy that prescribes to play as follows: play σ for the first $n-1$ stages and then, if $h^n = \bar{h}^n$, play j_ω, while if $h^n \neq \bar{h}^n$ then play an optimal mixed action in the matrix game A_{s^n}; and finally, play a δ-optimal strategy afterwards. Note that

$$\mathbb{P}_{s^1\pi\sigma^\delta}(\bar{h}^n) = \mathbb{P}_{s^1\pi\sigma}(\bar{h}^n) > 0.$$

Since we have chosen a shortest history $\bar{h}^n$ with the above property, by the definitions of X^* and Y^* we have

$$\mathbb{E}_{s^1\pi\sigma^\delta}\left(v_{s^n}\right) = v_{s^1},$$

and by the used mixed actions at stage n

$$\mathbb{E}_{s^1\pi\sigma^\delta}\left(v_{s^{n+1}}\right) \leq \mathbb{E}_{s^1\pi\sigma^\delta}\left(v_{s^n}\right) - \mathbb{P}_{s^1\pi\sigma^\delta}(\bar{h}^n)\cdot\tau.$$

From stage $n+1$, player 2 plays a δ-optimal strategy, so the choice of δ yields

$$\begin{aligned}\gamma(s^1,\pi,\sigma^\delta) &\leq \mathbb{E}_{s^1\pi\sigma^\delta}\left(v_{s^{n+1}}\right) + \delta \leq \mathbb{E}_{s^1\pi\sigma^\delta}\left(v_{s^n}\right) - \mathbb{P}_{s^1\pi\sigma^\delta}(\bar{h}^n)\cdot\tau + \delta \\ &= v_{s^1} - \mathbb{P}_{s^1\pi\sigma}(\bar{h}^n)\cdot\tau + \delta < v_{s^1},\end{aligned}$$

which contradicts the optimality of π. ∎

Based on the fact that any optimal strategy of player 1 in Γ guarantees the value v and, in view of the previous lemma, it only prescribes mixed actions in X^*_s, if the play is in state s, against any strategy of player 2 in Σ^*, we show that player 1 can guarantee at least v in the restricted game Γ^*. On the other hand, as discussed in (1.9), player 1 cannot guarantee more than the limit of the β-discounted values in Γ^*.

LEMMA 2.3. *Suppose that player* 1 *has an optimal strategy* $\pi \in \Pi$. *Then*

$$v_s \leq \sup_{\pi^*\in\Pi^*}\inf_{\sigma^*\in\Sigma^*}\gamma(s,\pi^*,\sigma^*) \leq v^{*1}_s \qquad \forall s \in S.$$

Proof. The second inequality follows from (1.9), so we only have to show the first one. For $\alpha,\omega \in S$ let

$$\bar{H}(\alpha,\omega) := \{h \in H(\alpha,\omega) |\, \mathbb{P}_{\alpha\pi\sigma^*}(h) > 0 \text{ for some } \sigma^* \in \Sigma^*\}.$$

Take an arbitrary $x \in X^*$. Using Lemma 2.2 we may define a strategy $\pi^* \in \Pi^*$ as follows: for $h \in H(\alpha,\omega)$ let

$$\pi^*(h) := \begin{cases} \pi(h) & \text{if } h \in \bar{H}(\alpha,\omega), \\ x_\omega & \text{otherwise.}\end{cases}$$

Then, by the optimality of π and by the definition of π^*, we have

$$v_s \leq \gamma(s, \pi, \sigma^*) = \gamma(s, \pi^*, \sigma^*) \qquad \forall \sigma^* \in \Sigma^*,\ \forall s \in S,$$

which implies the first inequality. □

The next result shows the effectiveness of the β-discounted optimal strategies in the restricted game Γ^*.

LEMMA 2.4. *Let $\varepsilon > 0$. For $\beta \in \mathcal{B}$, let $x^\beta \in X^*$ be a β-discounted optimal strategy of player 1 in Γ^*, and let $y \in Y^*$. Suppose that $E \subset S$ is a closed set of states with respect to (x^β, y) for all $\beta \in \mathcal{B}$. Then, for large $\beta \in \mathcal{B}$,*

$$\gamma(s, x^\beta, y) \geq \min_{t \in E} v_t^{*1} - \varepsilon \qquad \forall s \in E.$$

Proof. Using inequality (1.7) for Γ^* we have

$$(1-\beta)\, r(x^\beta, y) + \beta\, P_{x^\beta y}\, v^{*\beta} \geq v^{*\beta} \qquad \forall \beta \in \mathcal{B}.$$

By (1.1), multiplying this inequality with $Q_{x^\beta y}$ yields

$$Q_{x^\beta y}\, r(x^\beta, y) \geq Q_{x^\beta y}\, v^{*\beta} \qquad \forall \beta \in \mathcal{B}.$$

The closedness of E implies that, for any $s \in E$, if $q(t|s, x^\beta, y) > 0$ then $t \in E$. Hence for all $s \in E$ and for large $\beta \in \mathcal{B}$, using (1.2), we have

$$\begin{aligned} \gamma(s, x^\beta, y) &= \sum_{t \in E} q(t|s, x^\beta, y)\, r(t, x_t^\beta, y_t) \geq \sum_{t \in E} q(t|s, x^\beta, y)\, v_t^{*\beta} \\ &\geq \sum_{t \in E} q(t|s, x^\beta, y)(v_t^{*1} - \varepsilon) \geq \min_{t \in E} v_t^{*1} - \varepsilon, \end{aligned}$$

so the proof is complete. □

Next we discuss some properties of stationary strategies belonging to X^* or to $\mathrm{Relint}(X^*)$, where $\mathrm{Relint}(X^*)$ stands for the relative interior of the polytope X^* and is defined as the set of points in X^* which can be written as a convex combination of all the extreme points of X^* with only strictly positive coefficients.

LEMMA 2.5. *Let $x \in X^*$ and $y \in Y$. Suppose E is an ergodic set with respect to (x, y). Then $v_s = v_t$ for all $s, t \in E$. Furthermore, if $x \in \mathrm{Relint}(X^*)$, then necessarily $y_s \in Y_s^*$ for all $s \in E$.*

Proof. By $x \in X^*$ we obtain

$$\sum_{t \in E} p(t|s, x_s, y_s)\, v_t \geq v_s \qquad \forall s \in E.$$

Let $\bar{E} := \{s \in E |\, v_s = \max_{t \in E} v_t\}$. The above inequalities imply that $\bar{E}$ is a closed set of states for (x, y), so since E is an ergodic set for (x, y) (minimal closed set of states), we have $\bar{E} = E$. Therefore, $v_s = v_t =: v_E$ for all $s, t \in E$.

Now suppose that $x \in \mathrm{Relint}(X^*)$. Then $(\bar{x}_s, y_s)$ only induces transitions to states in E for any $\bar{x}_s \in X_s^*$, $s \in E$; hence

$$\sum_{t \in S} p(t|s, \bar{x}_s, y_s)\, v_t = \sum_{t \in E} p(t|s, \bar{x}_s, y_s)\, v_E = v_E = v_s \quad \forall \bar{x}_s \in X_s^*,\ \forall s \in E,$$

which implies that $y_s \in Y_s^*$ for all $s \in E$. □

An important property of convex combinations of stationary strategies is stated in the next lemma.

LEMMA 2.6. *For $\tau \in (0,1)$, $x^1, x^2 \in X$ let $x^\tau := \tau x^1 + (1-\tau)x^2$. Suppose that E is an ergodic set with respect to (x^τ, y) for some $y \in Y$. Let $\varepsilon > 0$ and $d \in \mathbb{R}$. If*

$$\gamma(s, x^1, y) \geq d \qquad \forall s \in E,$$

then for sufficiently large τ

$$\gamma(s, x^\tau, y) \geq d - \varepsilon \qquad \forall s \in E.$$

Proof. Let $\delta \in (0,1)$. Since

$$\gamma(s, x^1, y) \geq d \qquad \forall s \in E,$$

there exists a K^δ satisfying

$$\mathbb{E}_{sx^1y}(R_N) \geq d - \delta \qquad \forall N \geq K^\delta,\ \forall s \in E,$$

where R_N denotes the average payoff up to stage N. Choose $\tau \in (0,1)$ such that

$$\tau^{K^\delta} \geq 1 - \delta.$$

The strategy x^τ can be interpreted as playing x^1 with probability τ and x^2 with probability $1-\tau$ at each stage, so the last inequality means that x^1 will be played at each K^δ consecutive stages with probability at least $1-\delta$. Hence, with probability at least $1-\delta$, the expected average of the payoffs will be at least $d-\delta$ for any K^δ consecutive stages. Let r denote the smallest payoff in the game. Then if δ is small, by the law of large numbers we have

$$\gamma(s, x^\tau, y) \geq (1-\delta)(d-\delta) + \delta r \geq d - \varepsilon \qquad \forall s \in E,$$

so the proof is complete. □

The next lemma will enable us to construct Markov optimal strategies from stationary ε-optimal strategies which prescribe optimal mixed actions in the matrix games A_s, $s \in S$ (cf. (1.6)). Here we present a short proof, which uses some arguments of Bewley and Kohlberg [1978] on so-called irreducible games.

LEMMA 2.7. *Suppose that for all $\varepsilon > 0$ player 1 has a stationary ε-optimal strategy $x^\varepsilon \in X^*$ in Γ. Then player 1 also has a Markov optimal strategy f in Γ.*

Proof. Consider the restricted game $\Gamma^*(1)$, derived from Γ, where player 1 is restricted to use strategies that only prescribe mixed actions in X_s^*, if the play is in state $s \in S$. As before, Π^* will denote the set of these strategies for player 1. (Note that here only player 1 is restricted, in contrast with the game Γ^*, where both players have a restriction.) Let $v^{*\beta}(1)$ denote the β-discounted value in $\Gamma^*(1)$ and let $v^{*1}(1) := \lim_{\beta \uparrow 1} v^{*\beta}(1)$. Let $v^{*N}(1)$ denote the value of the N-stage game $\Gamma^{*N}(1)$, and let f^N be an N-stage Markov optimal strategy in $\Gamma^{*N}(1)$. Using the assumption that $x^\varepsilon \in X^*$ is ε-optimal in Γ for all $\varepsilon > 0$ and using (1.9) and (1.8), we obtain

$$v_s \leq \sup_{\pi \in \Pi^*} \inf_{\sigma \in \Sigma} \gamma(s, \pi, \sigma) \leq v_s^{*1}(1) = \lim_{N \to \infty} v_s^{*N}(1) \qquad \forall s \in S. \tag{2.1}$$

Let f be the Markov strategy of player 1 which prescribes to play as follows: at stage 1, play f^1; at the next two stages, play f^2; at the next three stages, play f^3; and so

on. We show that f is optimal. Let s^1 be the initial state and let s^N, $N \geq 2$, denote the state for the first stage when f^N is to be played. Take an arbitrary $\sigma \in \Sigma$. Notice that $f \in \Pi^*$, hence by the definition of X^*,

$$\mathbb{E}_{s^1 f \sigma}(v_{s^N}) \geq v_{s^1} \qquad \forall N \in \mathbb{N}.$$

Thus using the N-stage optimality of f^N and (2.1), for any $\delta > 0$ if N is large, then

$$\mathbb{E}_{s^1 f \sigma}(R^N) \geq \mathbb{E}_{s^1 f \sigma}\left(v^{*N}_{s^N}(1)\right) \geq \mathbb{E}_{s^1 f \sigma}\left(v_{s^N}\right) - \delta \geq v_{s^1} - \delta, \tag{2.2}$$

where R^N denotes the average payoff for those N consecutive stages when f^N is played. Let $\phi(K)$ be such that $f^{\phi(K)}$ is to be played at stage K. Observe that

$$\lim_{K\to\infty} \left[\frac{\sum_{N<\phi(K)} N}{K}\right] = 1, \qquad \lim_{K\to\infty} \left[\frac{K - \sum_{N<\phi(K)} N}{K}\right] = 0,$$

so if R_K denotes the average payoff up to stage K and r denotes the smallest payoff in the game, then (2.2) gives

$$\begin{aligned}
\gamma(s^1, f, \sigma) &= \liminf_{K\to\infty} \mathbb{E}_{s^1 f \sigma}\left(R_K\right) \\
&\geq \liminf_{K\to\infty} \mathbb{E}_{s^1 f \sigma}\left(\frac{\sum_{N<\phi(K)} N \cdot R^N + \left[K - \sum_{N<\phi(K)} N\right] \cdot r}{K}\right) \\
&= \liminf_{K\to\infty} \frac{\sum_{N<\phi(K)} N \cdot \mathbb{E}_{s^1 f \sigma}\left(R^N\right)}{K} \\
&\geq v_{s^1},
\end{aligned}$$

which implies that f is optimal in Γ. □

Now we are ready to prove Theorem 2.1.

Proof of Theorem 2.1. We show the existence of stationary ε-optimal strategies for all $\varepsilon > 0$, and then the existence of Markov optimal strategies follows from Lemma 2.7.

For $\beta \in \mathcal{B}$, let $x^\beta \in X^*$ be a β-discounted optimal strategy of player 1 in Γ^* and let $x \in \text{Relint}(X^*)$. For all $\tau \in (0,1)$ and $\beta \in \mathcal{B}$ let

$$x^{\tau\beta} := \tau x^\beta + (1-\tau)x.$$

By the convexity of X^* and by $x \in \text{Relint}(X^*)$ we have $x^{\tau\beta} \in \text{Relint}(X^*)$ for all $\tau \in (0,1)$ and $\beta \in \mathcal{B}$.

We show that, for any $\varepsilon > 0$, for large $\tau \in (0,1)$ and for large $\beta \in \mathcal{B}$ the strategy $x^{\tau\beta}$ is ε-optimal. Let $\varepsilon > 0$. Since against a stationary strategy there always exists a pure stationary best reply, and there are only finitely many pure stationary strategies, it suffices to show that, for all $j \in J$, if $\tau \in (0,1)$ and $\beta \in \mathcal{B}$ are large, then

$$\gamma(x^{\tau\beta}, j) \geq v - \varepsilon\, 1_z,$$

where $1_z = (1, \ldots, 1) \in \mathbb{R}^z$. Take a $j \in J$ and let $E \subset S$ be an arbitrary ergodic set with respect to $(x^{\tau\beta}, j)$. We start with showing that for large $\tau \in (0,1), \beta \in \mathcal{B}$ we have

$$\gamma(s, x^{\tau\beta}, j) \geq v_s - \varepsilon \qquad \forall s \in E. \tag{2.3}$$

Since $x^{\tau\beta} \in \text{Relint}(X^*)$, by Lemma 2.5 we obtain $v_s = v_t := v_E$ for all $s, t \in E$ and $j_s \in J_s^*$ for all $s \in E$. Let $j_s^* := j_s$ for all $s \in E$ and let $j_s^* \in J_s^*$ for all $s \notin E$; so $j^* \in J^*$. By the definition of $x^{\tau\beta}$ and by the properties of $\mathcal{B}$, the set of states E is closed with respect to (x^β, j) for all $\beta \in \mathcal{B}$, so with respect to (x^β, j^*) for all $\beta \in \mathcal{B}$ as well. Thus, applying Lemma 2.4 for Γ^* and using Lemma 2.3 yields that for large $\beta \in \mathcal{B}$

$$\gamma(s, x^\beta, j) = \gamma(s, x^\beta, j^*) \geq \min_{t \in E} v_t^{*1} - \frac{1}{2}\varepsilon \geq \min_{t \in E} v_t - \frac{1}{2}\varepsilon = v_E - \frac{1}{2}\varepsilon \qquad \forall s \in E.$$

Now Lemma 2.6 yields that for large $\tau \in (0,1)$ and for large $\beta \in \mathcal{B}$

$$\gamma(s, x^{\tau\beta}, j) \geq v_E - \varepsilon = v_s - \varepsilon \qquad \forall s \in E,$$

which proves (2.3).

Using that $x^{\tau\beta} \in X^*$ we have

$$P_{x^{\tau\beta} j}\, v \geq v,$$

therefore

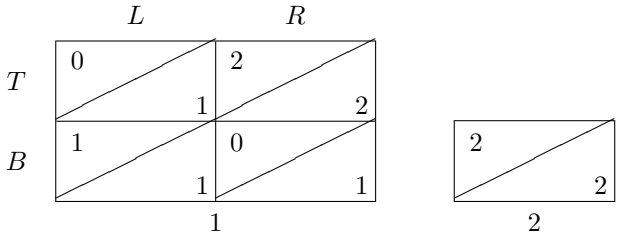
$$Q_{x^{\tau\beta} j}\, v \geq v.$$

For any $s \in S$, $q(t|s, x^{\tau\beta}, j) > 0$ implies that $t \in E$ for some ergodic set E with respect to $(x^{\tau\beta}, j)$; hence by (1.3) and (2.3), for large $\tau \in (0,1)$ and $\beta \in \mathcal{B}$, we obtain

$$\gamma(x^{\tau\beta}, j) = Q_{x^{\tau\beta} j}\, \gamma(x^{\tau\beta}, j) \geq Q_{x^{\tau\beta} j}\, (v - \varepsilon\, 1_z) = Q_{x^{\tau\beta} j}\, v - \varepsilon\, 1_z \geq v - \varepsilon\, 1_z,$$

which completes the proof. □

Example 1.



Here player 1 chooses rows and player 2 chooses columns. In each entry, the corresponding payoff is placed in the upper-left corner, while the transition is placed in the bottom-right corner. In this game each transition is represented by the number of the state to which transition should occur with probability 1. Notice that state 2 is absorbing. The value of this game is $v = (1, 2)$. It is not hard to show that there are optimal strategies for player 1 (later we will construct optimal Markov strategies). Following the construction for stationary ε-optimal strategies, we have $X^* = X$, $Y_1^* = \{(1,0)\}$, $Y_2^* = \{(1)\}$. Now the β-discounted optimal strategy of player 1 in Γ^* is $x^\beta = ((0,1),(1))$ for all $\beta \in (0,1)$. Take a strategy $x \in \text{Relint}(X^*)$, for example, $x = ((\frac{1}{2}, \frac{1}{2}), (1))$. Then for $\tau, \beta \in (0,1)$,

$$x^{\tau\beta} = \tau \cdot x^\beta + (1-\tau) \cdot x = \left(\left(\frac{1}{2} - \frac{1}{2}\tau, \frac{1}{2} + \frac{1}{2}\tau \right), (1) \right),$$

so $x^{\tau\beta}$ is ε-optimal for large τ and β (the strategies $((p,1-p),(1))$ are ε-optimal for $p \in (0,\varepsilon]$). Note that player 1 has no stationary optimal strategy in this game.

Also, a Markov optimal strategy can be constructed as in Lemma 2.7. In this game $X = X^*$, hence the restricted game $\Gamma^*(1)$ is just the original game Γ. The one-stage Markov optimal strategy and the one-stage value are

$$f^1 = \left(\left(\frac{1}{3},\frac{2}{3}\right),(1)\right), \qquad v^1 = v^{*1}(1) = \left(\frac{2}{3},2\right);$$

the two-stage Markov optimal strategy and the two-stage value are

$$f^2 = \left(\left(\left(\frac{3}{13},\frac{10}{13}\right),(1)\right);\left(\left(\frac{1}{3},\frac{2}{3}\right),(1)\right)\right), \qquad v^2 = v^{*2}(1) = \left(\frac{28}{39},2\right);$$

and so on. So, as shown before, the Markov strategy f which prescribes to play f^1 at the first stage, then f^2 at the next two stages, f^3 at the next three stages, and so on, is optimal.

3. Nonimproving strategies. It is in the spirit of zero-sum games to evaluate a strategy π of player 1 by the reward $\phi(\pi)$ it guarantees against any strategy of the opponent. For a strategy $\pi \in \Pi$ let

$$\phi_s(\pi) := \inf_{\sigma} \gamma(s,\pi,\sigma) \quad \forall s \in S, \qquad \phi(\pi) := (\phi_s(\pi))_{s\in S}\,.$$

Using this evaluation ϕ we may naturally define the relation "ε-better" between strategies of player 1. A strategy π^1 is called ε-better than π^2 if $\phi_s(\pi^1) \geq \phi_s(\pi^2) - \varepsilon$ holds for all $s \in S$. 0-better strategies will simply be called better. We will call a strategy π nonimproving if for any state $s \in S$ and for any history $h \in H(\cdot,s)$ we have

$$\phi_s(\pi) \geq \phi_s(\pi[h]).$$

Intuitively, a nonimproving strategy, for any state, cannot guarantee a larger reward conditional on any past history than initially. Obviously, all stationary strategies are nonimproving strategies. Also, optimal strategies are always nonimproving, since they guarantee the value, and no higher reward can be guaranteed.

In this section we will indicate how the following result, which is a generalization of Theorem 2.1, can be shown by using similar techniques as in section 2.

THEOREM 3.1. *For any nonimproving strategy, for any $\varepsilon > 0$, there exists an ε-better stationary strategy and a better Markov strategy as well.*

First we focus on the proof for the existence of ε-better stationary strategies, $\varepsilon > 0$, and afterwards we explain how the existence of a better Markov strategy will follow. Let π denote a fixed nonimproving strategy and let

$$w := \phi(\pi).$$

For $s \in S$ let

$$B_s := \left[\sum_{t\in S} p(t|s,i_s,j_s)\, w_t\right]_{i_s\in I_s,\, j_s\in J_s}, \qquad W_s := \mathrm{Val}\,(B_s),$$

where Val stands for the matrix game value. By using the nonimprovingness of π we obtain

$$\begin{aligned} w_s = \phi_s(\pi) &\leq \sum_{t\in S}\sum_{i_s\in I_s} \pi(s)(i_s)\, p(t|s,i_s,j_s)\cdot \phi_s(\pi[s,i_s,j_s,t]) \\ &\leq \sum_{t\in S}\sum_{i_s\in I_s} \pi(s)(i_s)\, p(t|s,i_s,j_s)\cdot w_t = \sum_{t\in S} p(t|s,\pi(s),j_s)\, w_t \qquad \forall j_s \in J_s,\ \forall s \in S, \end{aligned}$$

hence

$$w_s \le W_s = \mathrm{Val}\,(B_s) \qquad \forall s \in S. \tag{3.1}$$

This is the counterpart of (1.5), however, for w equality does not hold as for the value v, which causes some additional difficulties. We will define a restricted game here as well, but this restricted game will only be defined on a set of states s where $w_s = W_s$, so that we can use similar arguments as in section 2. Let

$$\tilde{X}_s := \left\{ x_s \in X_s | \sum_{t \in S} p(t|s, x_s, y_s)\, w_t \ge w_s \ \ \forall y_s \in Y_s \right\}, \qquad \tilde{X} := \times_{s \in S} \tilde{X}_s,$$

so the set $\tilde{X}_s$, which is a polytope, is the set of mixed actions of player 1 in state s which assure that after transition w will not decrease in expectation. The inequalities (3.1) imply that, for any state $s \in S$, all optimal mixed actions of player 1 in the matrix game B_s belong to $\tilde{X}_s$, which also means that the sets $\tilde{X}_s$ are nonempty.

Fix an arbitrary $x \in \mathrm{Relint}(\tilde{X})$. For a pure stationary strategy $j \in J$ let $R(j)$ denote the set of recurrent sets with respect to (x, j). Let

$$S' := \cup_{j \in J} R(j).$$

For $s \in S'$ let

$$J'_s := \cup_{j \in J,\, s \in R(j)} \{j_s\}, \quad Y'_s := \mathrm{conv}(J'_s), \quad Y' := \times_{s \in S'} Y'_s,$$

where conv stands for the convex hull of a set. Notice that the sets $R(j), S', J'_s, Y'_s, Y'$ are independent of the choice of $x \in \mathrm{Relint}(\tilde{X})$ and also that the sets Y'_s are nonempty polytopes. One can verify that all states $s \in S'$ are recurrent with respect to (x, y), if $y \in Y$ satisfies $y_s \in \mathrm{Relint}(Y'_s)$ for all $s \in S'$. If E is an ergodic set with respect to (x, y) with $y_s \in \mathrm{Relint}(Y'_s)$ for all $s \in S'$, then, as in Lemma 2.5, one can show that $w_s = w_t$ for all $s, t \in E$. Since $x \in \mathrm{Relint}(\tilde{X})$, this also yields that $w_s = W_s$ for all $s \in S'$, so w_s has a similar property as v_s in (1.5). The sets S' and Y' also have the property that, for any $y \in Y$, if $s \in S$ is recurrent with respect to (x, y), then $s \in S'$ and $y_s \in Y'_s$. Let

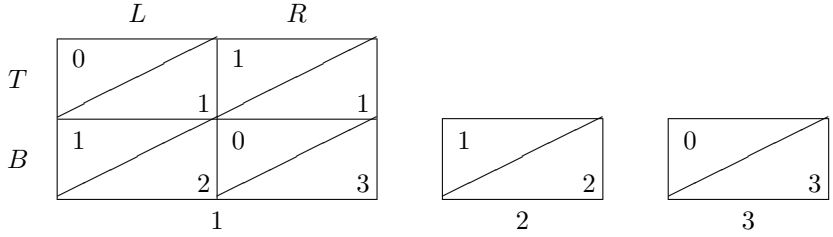
$$X' := \times_{s \in S'} \tilde{X}_s.$$

Let Γ' be the restricted game, derived from Γ, where the state space is S' and the players are restricted to using strategies that only prescribe mixed actions in X'_s and Y'_s, respectively, if the play is in state $s \in S'$. Note that, by the above property of S' and Y', if player 1 uses mixed actions in $\mathrm{Relint}(X'_s)$, $s \in S'$, then whatever stationary strategy y player 2 uses, the play will eventually reach an ergodic set $E \subset S'$ in such a way that w does not decrease in expectation, and $y_s \in Y'_s$ for all $s \in E$, so intuitively the play will eventually proceed in Γ'. Now, using $w_s = W_s$ for all $s \in S'$, for the restricted game Γ', similar results can be shown as for the restricted game in section 2, which completes the proof for the existence of ε-better stationary strategies.

Now the existence of better Markov strategies can be shown along similar lines as the proof of Lemma 2.7. One has to define a restricted game $\Gamma'(1)$, derived from Γ, where player 1 is restricted to use strategies that only prescribe mixed actions in X'_s if the play is in state $s \in S$. Notice that $\Gamma'(1)$ is the counterpart of $\Gamma^*(1)$ defined in the proof of Lemma 7 and also that the above constructed ε-better stationary strategies

belong to X', hence player 1 may use these strategies in the restricted game $\Gamma'(1)$ as well. Now in the game $\Gamma'(1)$, analogous equalities and inequalities can be derived as for $\Gamma^*(1)$ in the proof of Lemma 2.7, but w has to be used instead of v, which leads to the conclusion that better Markov strategies indeed exist.

Example 2.



This example, known as the Big Match (cf. Gillette [1957], Blackwell and Ferguson [1968]), clarifies that, although optimality implies nonimprovingness, improving strategies are indispensable for achieving ε-optimality. The notation is the same as in Example 1. Notice that states 2 and 3 are absorbing. For initial state 1, the limiting average value is $v_1 = \frac{1}{2}$ and player 1 has neither optimal strategies nor stationary ε-optimal strategies for small $\varepsilon > 0$, but for any $N \in \mathbb{N}$ player 1 can guarantee $\frac{1}{2} - \frac{1}{2(N+1)}$ by playing the following strategy π^N: for any history h without absorption, if $k(h)$ denotes the number of stages where player 2 has chosen action R minus the number of stages where player 2 has chosen action L, player 1 has to play the mixed action

$$\pi^N(h) := \left(1 - \frac{1}{(k(h)+N+1)^2}, \frac{1}{(k(h)+N+1)^2}\right).$$

This strategy π^N is clearly improving, since for the history $h = (1, T, R, 1)$ we have $\pi^N[h] = \pi^{N+1}$. Note that, in fact, all strategies that are ε-optimal for small $\varepsilon > 0$ must be improving; otherwise, by Theorem 3.1, player 1 would have stationary ε-optimal strategies (and Markov optimal strategies as well).

4. Concluding remarks. Finally we discuss some consequences. For the sake of simplicity, we only focus on the results of section 2 here.

Remarks on the restricted game Γ^*. In Lemma 2.3 we showed that $v_s^{*1} \geq v_s$ for all $s \in S$. In fact, this is the only statement for which we needed the condition that player 1 has an optimal strategy. Therefore, if in a zero-sum game $v_s^{*1} \geq v_s$ holds for all $s \in S$, then stationary ε-optimal strategies, $\varepsilon > 0$, and Markov optimal strategies can be constructed exactly as in section 2. It also means that $v_s^{*1} \geq v_s$ for all $s \in S$ holds if and only if player 1 has an optimal strategy.

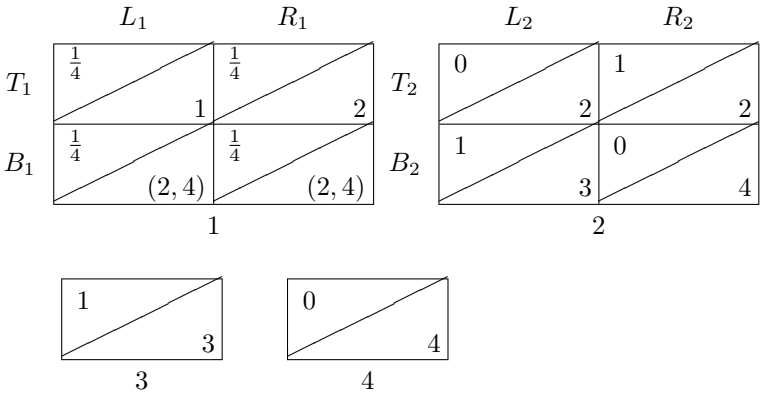
We also remark that, even if player 1 has an optimal strategy, one can find examples where $v_s^{*1} > v_s$ for some state s. However, if E is an ergodic set with respect to some $(x, y) \in \text{Relint}(X^*) \times \text{Relint}(Y^*)$, then there exists a state $s \in E$ such that $v_s^{*1} = v_E$ (recall that the value v is a constant on E by Lemma 2.5). To see this one can argue as follows. Suppose to the contrary that $v_s^{*1} \geq v_E + \mu$ for all $s \in E$, where $\mu > 0$. Let $x^{\tau\beta} \in \text{Relint}(X^*)$ be defined as in the proof of Theorem 2.1. Then Lemmas 2.4 and 2.6 imply that for large τ and β we have

$$\gamma(s, x^{\tau\beta}, j) \geq \min_{t \in E} v_t^{*1} - \frac{\mu}{2} \geq v_E + \frac{\mu}{2} \qquad \forall s \in E,\ \forall j \in J^*. \tag{4.1}$$

Here we used that there are only finitely many pure stationary strategies. Let player 1 play the strategy π^δ, $\delta > 0$, which prescribes to play as follows: play $x^{\tau\beta}$ as long as player 2 chooses actions in J_s^*, $s \in E$, and start playing a δ-optimal strategy as soon as player 2 chooses an action in $J_s \setminus J_s^*$ in some state $s \in E$. Note that if player 2 always chooses actions in J_s^*, $s \in E$, then (4.1) assures that the reward is at least $v_E + \frac{\mu}{2}$ (recall that against a stationary strategy there always exists a pure stationary best reply). On the other hand, if player 2 chooses an action in $J_s \setminus J_s^*$ in some state $s \in E$, then one can show that $x_s^{\tau\beta} \in \mathrm{Relint}(X_s^*)$ yields that the original value v increases in expectation by at least some $\nu > 0$, so if $\delta \in (0, \frac{\nu}{2})$, by the definition of π^δ, the reward is at least $v_E + \frac{\nu}{2}$ in this case. Therefore, π^δ, with $\delta \in (0, \frac{\nu}{2})$, guarantees a reward of at least $v_E + \frac{1}{2}\min(\mu, \nu) > v_E$, which contradicts the definition of the value. So we have shown that $v_s^{*1} = v_E$ holds for some state $s \in E$.

Optimal strategies for particular initial states. We briefly discuss a generalization of the results of section 2, which concerns strategies that are only optimal for particular initial states. Let $\tilde{S}$ denote the set of states for which player 1 has an optimal strategy. First note that in each stochastic game there always exists at least one initial state for which player 1 has optimal strategies (cf. Thuijsman and Vrieze [1991]), so the set $\tilde{S}$ is always nonempty. Using similar techniques as in section 2, one can show that, for any $\varepsilon > 0$, player 1 has a strategy ξ^ε which for all initial states $\alpha \in \tilde{S}$ satisfies the following criteria: (i) ξ^ε is ε-optimal, (ii) ξ^ε is stationary until leaving $\tilde{S}$, (iii) there exist stationary best replies of player 2 against ξ^ε, (iv) the probability of ever leaving $\tilde{S}$ is zero with respect to $(\alpha, \xi^\varepsilon, \sigma)$, if σ is a best reply. The difference between this result and the corresponding result of section 2 is mainly due to the fact that stationary strategies are not effective in states outside $\tilde{S}$, so player 1 may have to start playing a behavior δ-optimal strategy if the play leaves $\tilde{S}$, for some $\delta > 0$. Furthermore, one can also show that player 1 has a strategy χ which for all initial states $\alpha \in \tilde{S}$ satisfies the following criteria: (v) χ is optimal, (vi) χ is Markov until leaving $\tilde{S}$, (vii) there exist Markov best replies of player 2 against χ, (viii) the probability of ever leaving $\tilde{S}$ is zero with respect to (α, χ, σ), if σ is a best reply. We remark here that Markov best replies do not necessarily exist against a Markov strategy, but a Markov strategy χ can be constructed so that (vii) holds.

Example 3.



This example clarifies the existence of such "almost stationary" ε-optimal strategies and "almost Markov" optimal strategies for initial states in $\tilde{S}$. The notation is the same as in Example 1 except for two "mixed" transition vectors in entries (B_1, L_1) and (B_1, R_1), which lead to state 2 with probability $\frac{1}{2}$ and to state 4 with probability

$\frac{1}{2}$. For the sake of simplicity, we only focus on the possible simplifications by "almost stationary" ε-optimal strategies. Notice that if the initial state is state 2, then this game reduces to Example 2. So here the value is $v = (\frac{1}{4}, \frac{1}{2}, 1, 0)$. As mentioned, for initial state 2, player 1 has no optimal strategy, so $\tilde{S} = \{1, 3, 4\}$. Since initial states $3, 4 \in \tilde{S}$ are trivial, we assume the initial state to be $1 \in \tilde{S}$. Consider the strategy ξ for player 1 which prescribes playing action T_1 as long as the play is in state 1, and as soon as the play visits state 2 then prescribes starting to play a behavior $\frac{1}{8}$-optimal strategy. This strategy ξ is optimal and clearly satisfies properties (i), (ii), (iii), and (iv). Note that switching to a behavior strategy when entering state 2 is crucial, because by stationary strategies player 1 could only guarantee 0 for initial state 2. Note also that the use of action B_1 would violate property (iv).

An alternative proof for Lemma 2.7. We wish to remark that, under the condition of Lemma 2.7, other Markov optimal strategies exist as well . Let ε_n be a positive sequence converging to zero. One can show that the Markov strategy which prescribes x^{ε_1} for the first N_1 stages, x^{ε_2} for the next N_2 stages, and so on, is optimal for a well-chosen increasing sequence N_n.

Subgame optimality. Note that the Markov strategy f, constructed in section 2, is "subgame optimal"; namely, the strategy $f[h]$ is optimal for any finite history h.

Alternative rewards and optimality. It is worthwhile to mention that sometimes other rewards are used to evaluate the long-term average payoffs. The most common rewards are the following ones:

$$\gamma^1(s,\pi,\sigma) = \mathbb{E}_{s\pi\sigma}\left(\liminf_{N\to\infty} R_N\right), \qquad \gamma^2(s,\pi,\sigma) = \liminf_{N\to\infty} \mathbb{E}_{s\pi\sigma}\left(R_N\right),$$

$$\gamma^3(s,\pi,\sigma) = \limsup_{N\to\infty} \mathbb{E}_{s\pi\sigma}\left(R_N\right), \qquad \gamma^4(s,\pi,\sigma) = \mathbb{E}_{s\pi\sigma}\left(\limsup_{N\to\infty} R_N\right),$$

where R_N is the random variable for the average payoff up to stage $N \in \mathbb{N}$. It holds that $\gamma^1 \le \gamma^2 \le \gamma^3 \le \gamma^4$. Notice that we have used $\gamma = \gamma^2$ so far. Mertens and Neyman [1981] showed that the value is the same for all these rewards. Optimality and ε-optimality can be defined with respect to any of these four rewards. Sometimes a fifth alternative is to require uniformity from the optimal strategy; i.e., π is uniformly optimal for state $s \in S$ if

$$\forall \delta > 0\ \exists N^\delta\colon\ \mathbb{E}_{s\pi\sigma}\left(R_N\right) \ge v_s - \delta \quad \forall N \ge N^\delta,\ \forall \sigma \in \Sigma.$$

The definition of uniform ε-optimality is similar.

Focussing only on section 2 again, we briefly examine the validity of the results for all these criteria. First notice that it makes no difference in our results in which way the strategy of player 1 is optimal. Furthermore, for stationary strategy pairs, all the above optimality criteria are known to be equivalent (for example, cf. Bewley and Kohlberg [1978]), so the simplifications by stationary strategies remain valid with respect to all these alternatives. For Markov strategies, however, it is somewhat different. Notice first that the Markov strategy constructed in section 2 is uniformly optimal (see the proof of Lemma 2.7). Since $\gamma^2 \le \gamma^3 \le \gamma^4$ we have that this Markov strategy is also optimal for rewards γ^3, γ^4. However, when player 1 has an optimal strategy, the existence of Markov optimal strategies for reward γ^1 is not straightforward, not even by using an approach as in the alternative proof for Lemma 2.7.

REFERENCES

T. Bewley and E. Kohlberg [1976], *The asymptotic theory of stochastic games*, Math. Oper. Res., 1, pp. 197–208.

T. Bewley and E. Kohlberg [1978], *On stochastic games with stationary optimal strategies*, Math. Oper. Res., 3, pp. 104–125.

D. Blackwell and T. S. Ferguson [1968], *The big match*, Ann. Math. Stat., 33, pp. 159–163.

J. L. Doob [1953], *Stochastic Processes*, Wiley, New York.

D. Gillette [1957], *Stochastic games with zero stop probabilities*, in Contributions to the Theory of Games III, M. Dresher, A. W. Tucker, and P. Wolfe, eds., Ann. of Math. Stud. 39, Princeton University Press, Princeton, NJ, pp. 179–187.

A. Hordijk, O. J. Vrieze, and G. L. Wanrooij [1983], *Semi-Markov strategies in stochastic games*, Internat. J. Game Theory, 12, pp. 81–89.

J. F. Mertens and A. Neyman [1981], *Stochastic games*, Internat. J. Game Theory, 10, pp. 53–66.

T. E. S. Raghavan and J. A. Filar [1991], *Algorithms for stochastic games*, Z. Oper. Res., 35, pp. 437–472.

L. S. Shapley [1953], *Stochastic games*, Proc. Nat. Acad. Sci. U.S.A, 39, pp. 1095–1100.

F. Thuijsman and O. J. Vrieze [1991], *Easy initial states in stochastic games*, in Stochastic Games and Related Topics, T. E. S. Raghavan, T. S. Ferguson, O. J. Vrieze, and T. Parthasarathy, eds., Kluwer, Dordrecht, the Netherlands, pp. 85–100.